Document Images & ML

A COLLABORATORY BETWEEN THE LIBRARY OF CONGRESS AND THE IMAGE ANALYSIS FOR ARCHIVAL DISCOVERY (AIDA) LAB AT THE UNIVERSITY OF NEBRASKA, LINCOLN, NE

Yi Liu, Chulwoo Pack, Leen-Kiat Soh, Elizabeth Lorang, August 22, 2019

Overview of Projects

- Project 1: Document Segmentation (Mike & Yi)
- Project 2: Document Type Classification (Mike & Yi)
- Project 3: Quality Assessment (Yi)
- Project 3.1: Figure/Graph Extraction from Document (Yi)
- Project 3.2: Text Extraction from Figure/Graph (Yi)
- Project 4.1: Subjective Quality Assessment (Yi) (Work In Progress)
- Project 4.2: Objective Quality Assessment (Yi)
- Project 5: Digitization Type Differentiation: Microfilm or Scanned (Yi)

Background | State-of-the-Art CNN models

- **Convolutional Neural Network** (CNN) Models (deep learning)
  - Classification [Dataset; Top-1 / Top-5]
    - 2014, VGG-16 (Classification) [ImageNet; 74.4% / 91.9%]
    - 2015, ResNet-50 (Classification) [ImageNet; 77.2% / 93.3%]
    - 2018, ResNeXt-101 (Classification) [ImageNet; 85.1% / 97.5%]
  - Segmentation [Dataset; Intersection-over-Union (IoU)]
    - 2015, U-net (Segmentation/Pixel-wise classification) [ISBI; 92.0%]

- So, we now know that CNNs achieve *remarkable* performances in both classification and segmentation tasks.
- ***What about document images then?***

Project 1: Document Segmentation

Objectives | Find and localize *Figure*/*Illustration*/*Cartoon* presented in an image

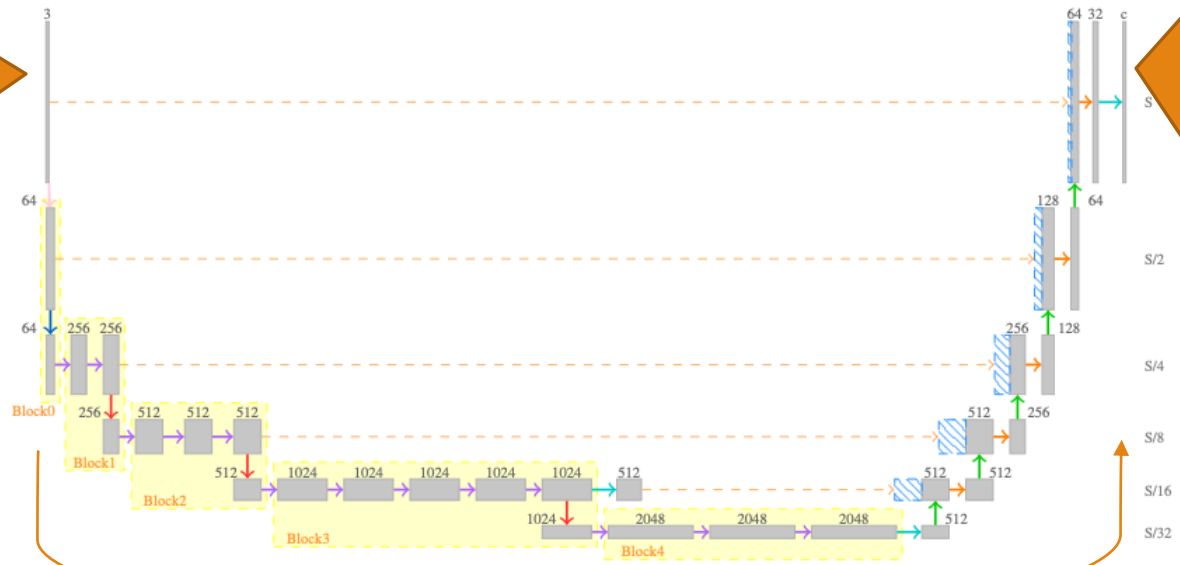
Applications | metadata generation, discover-/search-ability, visualization, etc.

Aida

Document Segmentation | Technical Details

- ***Training*** is a process of finding the optimal value weights between artificial neurons that minimizes a pre-defined ***loss*** function

Input

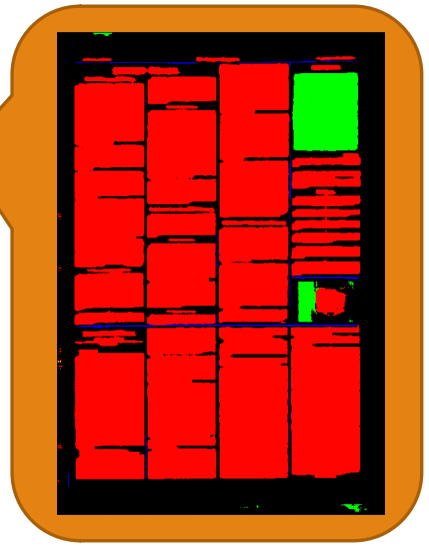


1. Convolution & Down-sampling:

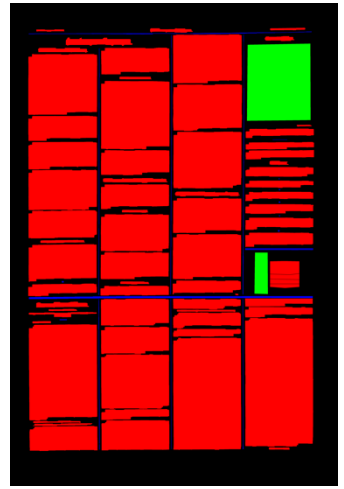
understand "***WHAT***" is present in the image (i.e., feature extraction)

2. Up-sampling:

understand "***WHERE***" it is present in the image



Prediction



Ground-truth

3. Calculate per-pixel loss
4. Update weights between neurons
5. Repeat the process

Aida

Document Segmentation | Dataset

Beyond Words

- Total of 2,635 image snippets from 1,562 pages (as of 7/24/2019)
  - 1,027 pages with single snippet
  - 512 pages with multiple snippets
- Issues
  - Inconsistency (Figure 1)
  - Imprecision (Figure 2)
  - Data imbalance (Figure 3)

Figure 1. Example of inconsistency. Note that there are more than one image snippets in the left image (i.e. input) while there is only a single annotation in the right ground-truth.

Figure 2. Example of imprecision. From left to right: (1) ground-truth (yellow: Photograph and black: background) and (2) original image. Note here that in the ground-truth, non-photograph-like (e.g., texts) components are included within the yellow rectangle region.

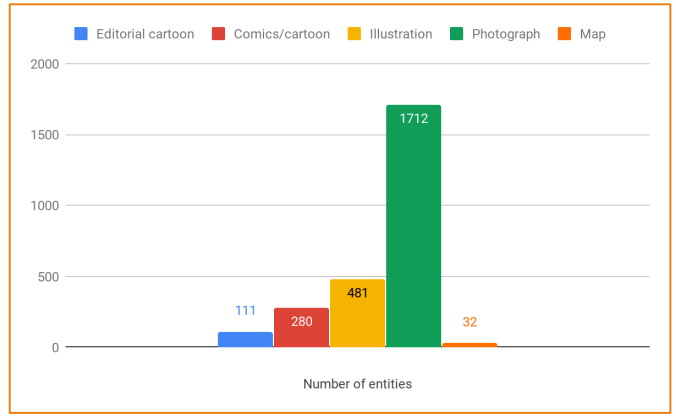


Figure 3. Number of snippets in Beyond Words. Note here the data imbalance

Document Segmentation | Dataset

European Historical Newspapers (ENP)

- Total of 57,339 image snippets in 500 pages
  - All pages have multiple snippets
- Issues
  - Data imbalance
    - Text: 43,780
    - Figure: 1,452
    - Line-separator: 11,896
    - Table: 221

Figure 4. Example of image (left) and ground-truth (right) from ENP dataset. In the ground-truth, each color represents the following components: (1) black: background, (2) red: text, (3) green: figure, (4) blue: line-separator, and (5) yellow: table.

Aida

Document Segmentation | Experimental Results

- A U-net model trained with ENP dataset shows better segmentation performance than that with Beyond Words in terms of pixelwise-accuracy and IoU score
  - IoU score is a commonly used metric to evaluate segmentation performance
  - The three issues—inconsistency, imprecision, and data imbalance—of Beyond Words dataset need to be improved for better use in training

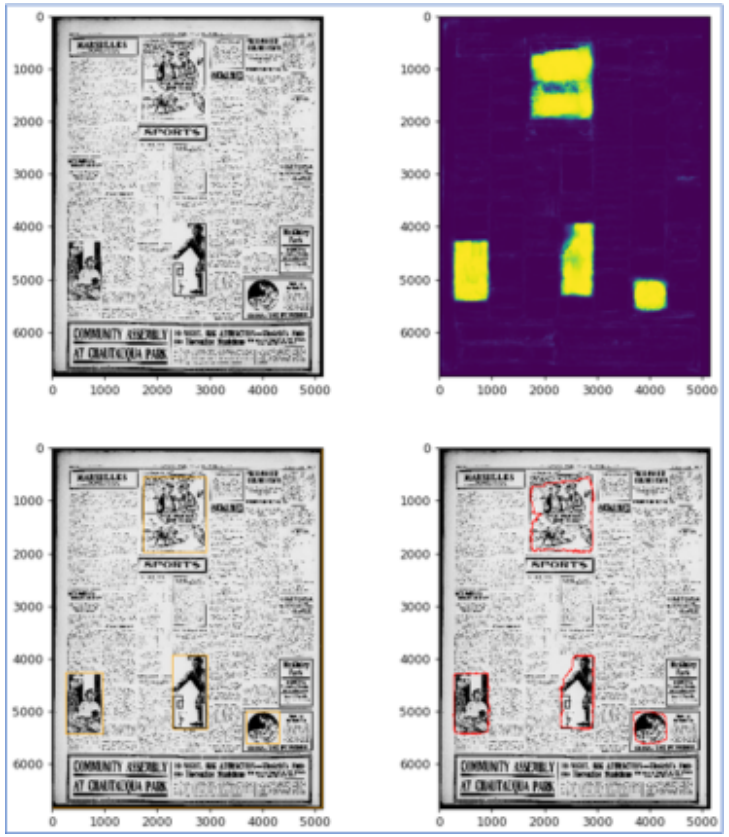
Model	train/eval size	Classes	Weighted training	Pre-processing (Normalization)	Best Score	
					Accuracy	mIoU
BW_1500_v1	1226/306	0: Background 1: Editorial cartoon 2: Comics/cartoon 3: Illustration 4: Photograph 5: Map	No	No	0.87	0.24
BW_1500_v2			Yes [10;22;20;18;8;22]		0.88	0.26
ENP_500_v1	385/96	0: Background 1: Text 2: Figure 3: Separator 4: Table	Yes [5;10;40;10;35]	No	0.88	0.64
ENP_500_v2				Yes	0.89	0.64
ENP_500_v3			**No**	**No**	**0.91**	**0.69**
ENP_500_v4				**Yes**	**0.91**	**0.69**

*Accuracy: Pixel-wise accuracy.
*mIoU: Average intersection over union.
*Normalization: Zero mean unit variance

- Assigning different weights per class to mitigate data imbalance did *not* show performance improvement
  - ***Future Work:*** Explore a different way of weighting strategy to mitigate a data imbalance problem

Aida

Document Segmentation | Potential Applications 1



- Enrich page-level metadata by cataloging the types of visual components presented on a page
- Enrich collection-level metadata as well
- Visualize figures' locations on a page

Figure 5. Segmentation result of ENP_500_v4 on Chronicling America image (sn92053240-19190805.jpg). Clockwise from top- left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability to belong to figure class are shown with brighter color.

Document Segmentation | Potential Applications 2

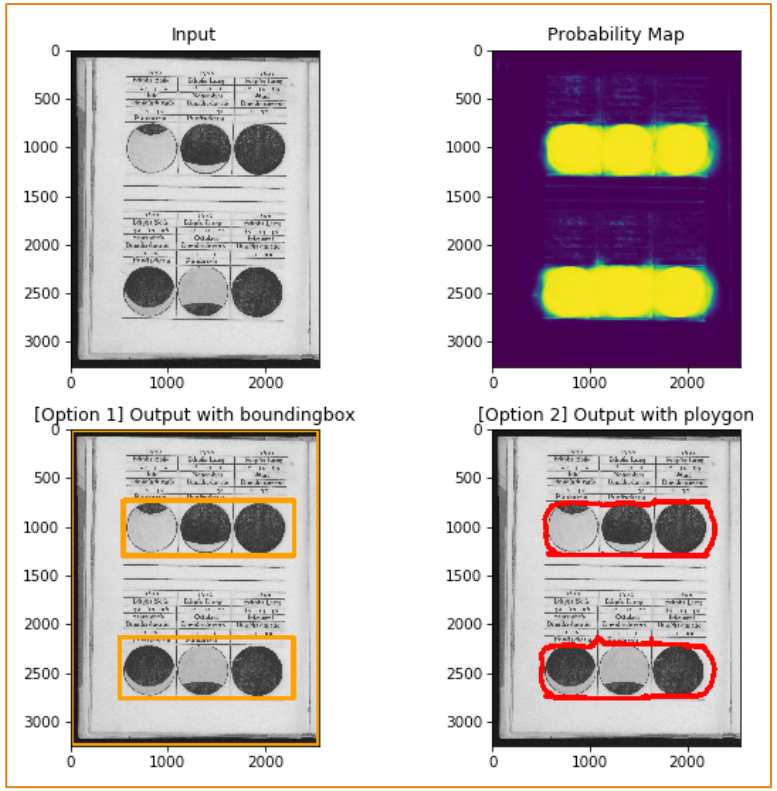
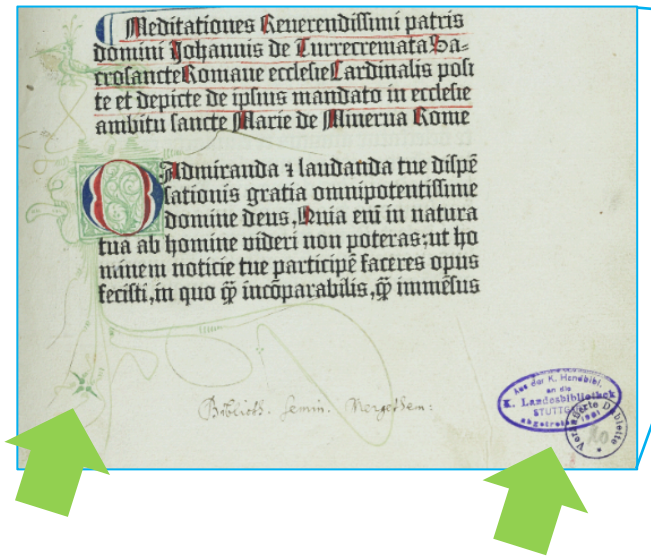


Figure 6. Successful segmentation result of ENP_500_v4 on book/printed material (https://www.loc.gov/resource/rbc0001.2013rosen0051/?sp=37).



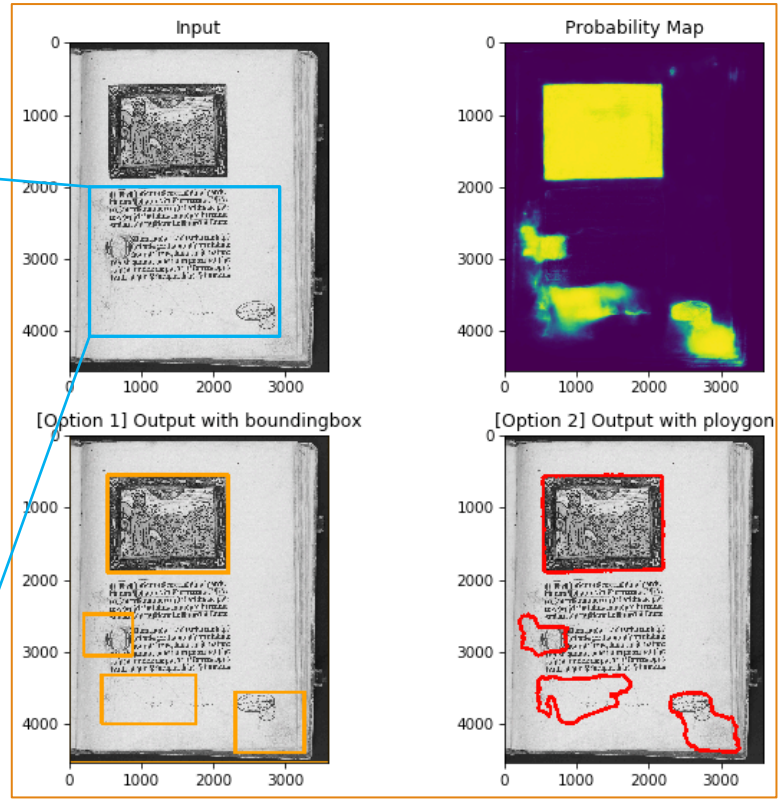


Figure 7. Failure segmentation result of ENP_500_v4 on book/printed material (https://cdn.loc.gov/service/rbc/rbc0001/2010/2010rosen0073/0005v.jpg). Note that there is light drawing or stamps (marked in green arrows) on the false positive regions.

Document Segmentation | Conclusions

- As a preliminary experiment, a state-of-the-art CNN model (i.e., U-net) shows promising segmentation performance on ENP document image dataset,
  - There is still room for improvement with more sophisticated training strategies (e.g., weighted training, augmentation, etc.)
- To make Beyond Words dataset more as a valuable training resource for machine learning researchers, we need to address the following issues:
  - Consistency
  - Precision of the coordinates of regions

Project 2: Document Type Classification

Objectives | (1) Classify a given image into one of *Handwritten*/*Typed*/*Mixed* type; (2) Classify a given image into one of *Scanned*/*Microfilmed*

Applications | metadata generation, discover-/search-ability, cataloging, etc.

Document Type Classification | Technical Details

*Note that we do not need up-sampling in this task, since **WHERE** is not our concern*

- A simple VGG-16 is used (Figure 8)
  - Afzal et al. reported that most of state-of-the-art CNN models yielded around 89% of accuracy on document image classification task
- **Transfer learning?**
  - Why don't we initialize our model's weights from a model that has been already trained on a large-scale data, such as *ImageNet* (about 14M images)?
  - ***Why?*** (1) training a model from the scratch (i.e., the value of weights between neurons are initialized to random number) takes too much time; (2) we have too small a dataset to train a model

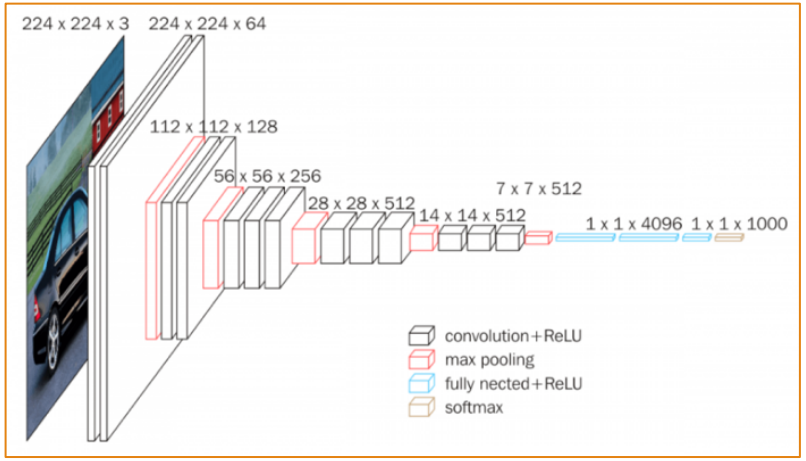


Figure 8. Architecture of original VGG-16. In our project, the last softmax layer is adjusted to have a shape of 3, which is the number of our target classes; handwritten, typed, and mixed

Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*(Vol. 1, pp. 883-888). IEEE.

Aida

Document Type Classification | Datasets

- We have two datasets:
  - Experiment 1: *RVL-CDIP* (400,000 document images with 16 different balanced classes); publicly available
  - Experiment 2: *suffrage_1002* (1,002 document images with 3 different balanced classes); manually compiled from ***By the People: Suffrage*** campaign (Table 1)

	handwritten	**typed**	**mixed**	Total
train	267	267	267	801
validation	33	33	33	99
test	33	33	33	99
Total	333	333	333	999

Table 1. Configuration of *suffrage_1002* dataset.

Figure 9. Example document images from each 16 different classes

Document Type Classification | Datasets

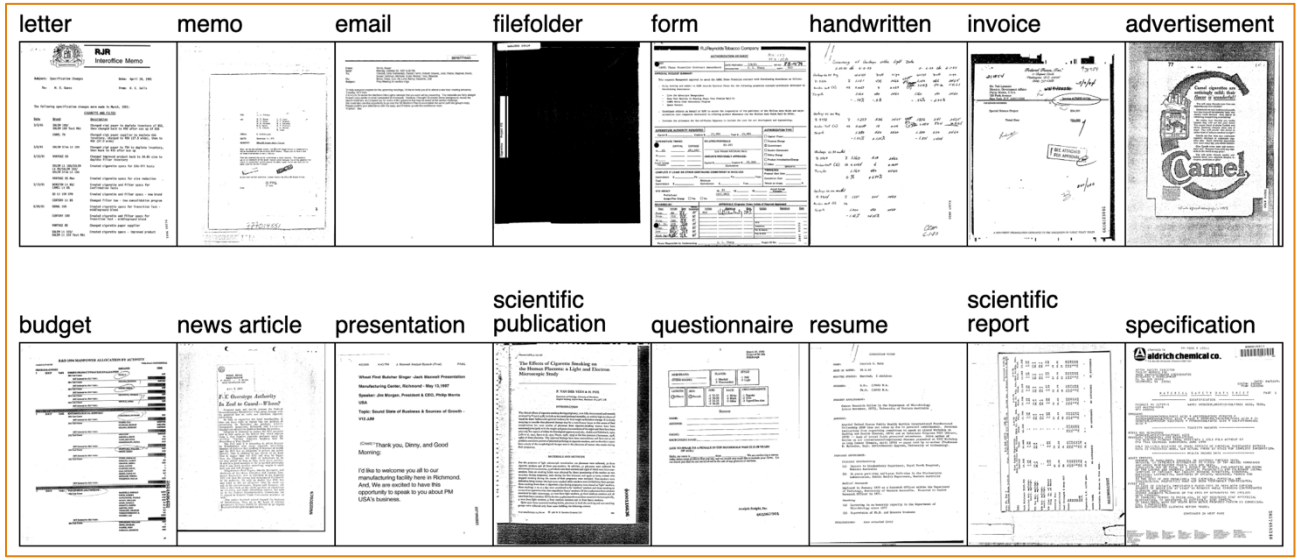


Figure 9. Example document images from each 16 different classes in *RVL_CDIP* dataset

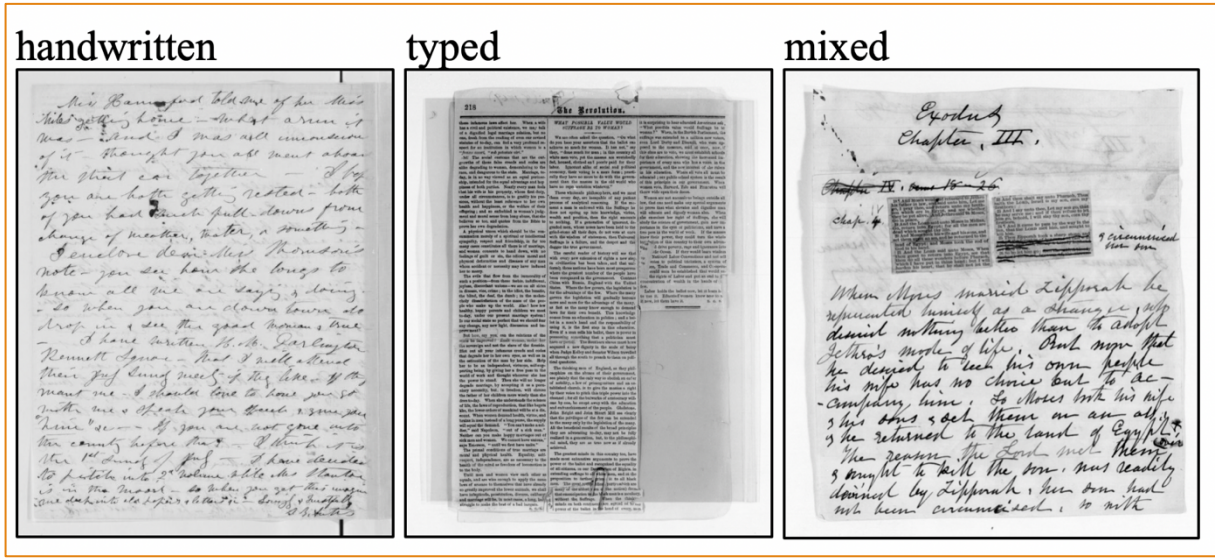


Figure 10. Example document images from each 3 different classes in *suffrage_1002* dataset

Document Type Classification | Experimental Results

Table 1. Precision, recall, and f1-score of *VGG-16* trained on *RVL_CDIP* dataset. The alphabetic labels are corresponding to the following labels: *letter*, *form*, *email*, ***handwritten***, *advertisement*, *scientific report*, *scientific publication*, *specification*, *file folder*, *news article*, *budget*, *invoice*, *presentation*, *questionnaire*, *resume*, and *memo*. Our class of interest, ***handwritten***, is bolded.

(unit: %)	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Avg
Precision	86	74	98	**89**	89	73	90	88	89	92	87	91	78	91	92	88	87
Recall	94	79	97	**96**	91	73	93	91	97	86	83	86	79	73	94	91	87
F1	86	77	97	**92**	90	73	91	90	93	89	85	88	79	81	93	90	87

Table 2. Precision, recall, and f1-score of *VGG-16* on *suffrage_1002* testing set.

(unit: %)	**handwritten**	**typed**	**mixed**	**Avg**
Precision	89	91	90	90
Recall	97	94	79	90
F1	93	93	84	90

- Experiment 1: We obtained a model trained on a large-scale document image dataset, *RVL-CDIP* with promising classification performance, as shown in Table 1
  - ***Implication***: Features learned from natural images (ImageNet) are general enough to apply to document images
  - Now we can utilize this model by retraining it with our own *suffrage_1002* dataset in Experiment 2
- Experiment 2: The retrained model shows even better classification performance, as shown in Table 2

Aida

Document Type Classification | Conclusions

- In both experiments, the state-of-the-art CNN model is capable of classifying document images with promising performance
  - ***Potential Applications***: help tagging an image type
- A main *challenge*: classifying a mixed type document image, as shown in Figure 11
  - ***Future Work:*** Perform a confidence level analysis to mitigate this problem
- ***Future Work:*** We expect that the classification performance can be further improved with a larger large-scale dataset

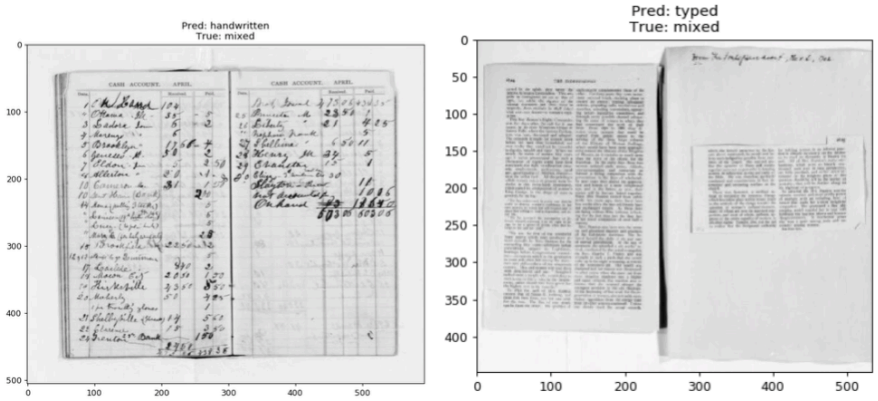


Figure 11. Failure prediction cases. On the left example, a typed region is relatively smaller than that of handwriting. On the right example, a handwriting region is relatively smaller than that of typing.

Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*(Vol. 1, pp. 883-888). IEEE.

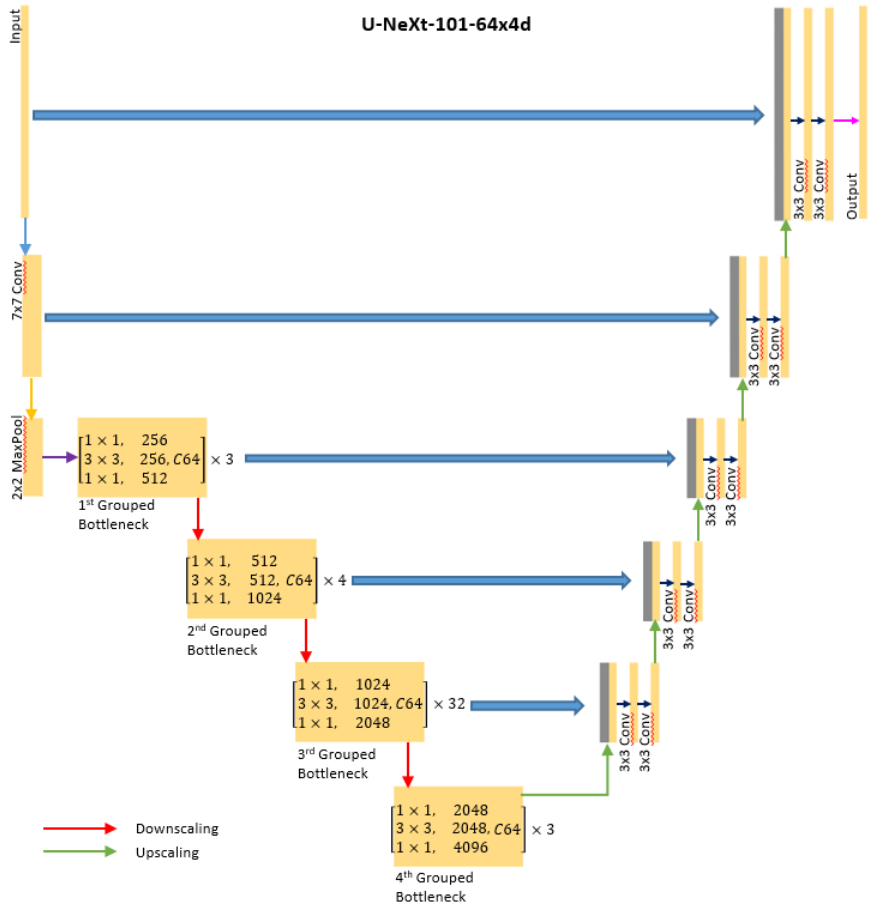
Project 3.1: Figure/Graph Extraction from Document

Objectives | Find and localize *Figure*/Graph in a document image

Applications | Graph retrieval, document segmentation based on content type

Aida

Figure/Graph Extraction from Document | Technical Details



An FCN (U-NeXt) is used

- U-NeXt combines ResNeXt and U-Net
  - ResNeXt101_64x4d
- Why ResNeXt101_64x4d?
  - Current state-of-art
  - Accessible pre-trained model
- **Transfer learning**
  - ResNeXt101_64x4d
  - Number of parameters:
    - 114.4 million → 32.8 million

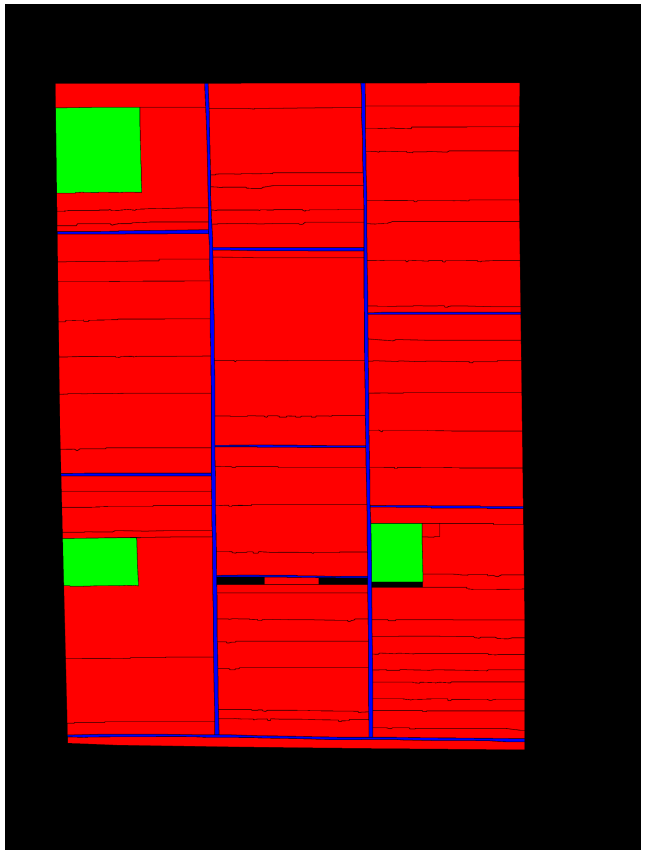
Aida

Figure/Graph Extraction from Document | Datasets

- **ENP collection**: European newspaper collection
  - A subset used for the International Conference on Document Analysis and Recognition competition
- **Beyond Word collection**: Transcribed collection
  - But cannot be used for training directly …
    - Problem 1: missing figures in ground-truth
    - Problem 2: inaccurate ground-truth

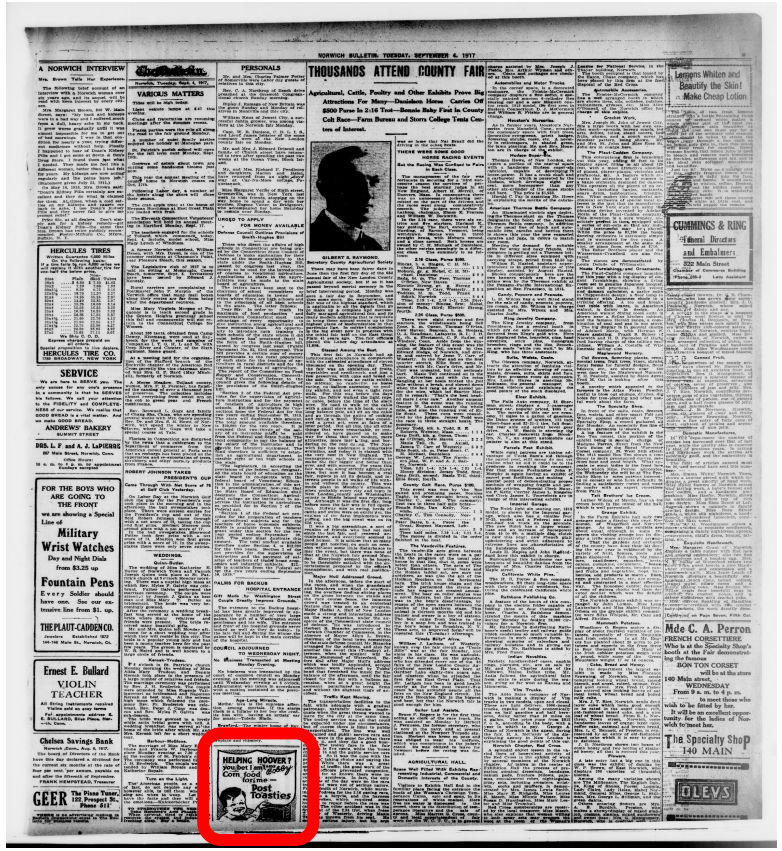
Figure/Graph Extraction from Document | Datasets: ENP

Document Image



Ground-truth

Figure/Graph Extraction from Document | Datasets: Beyond Words



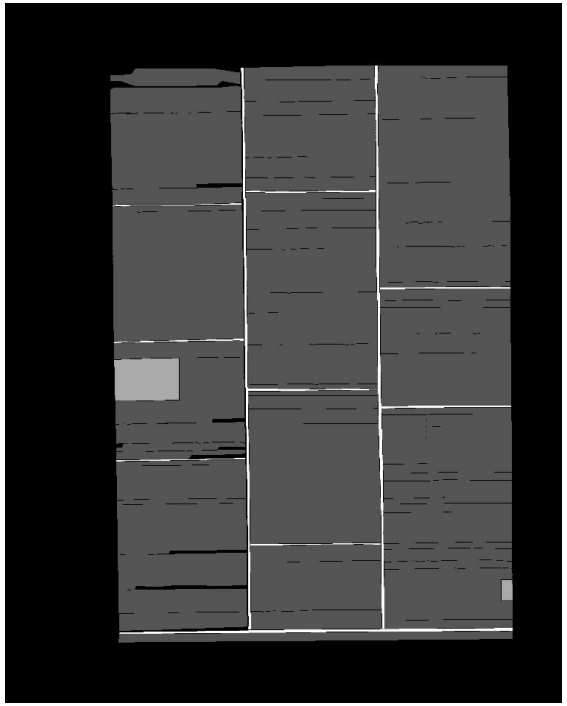
Document Image

Ground-truth

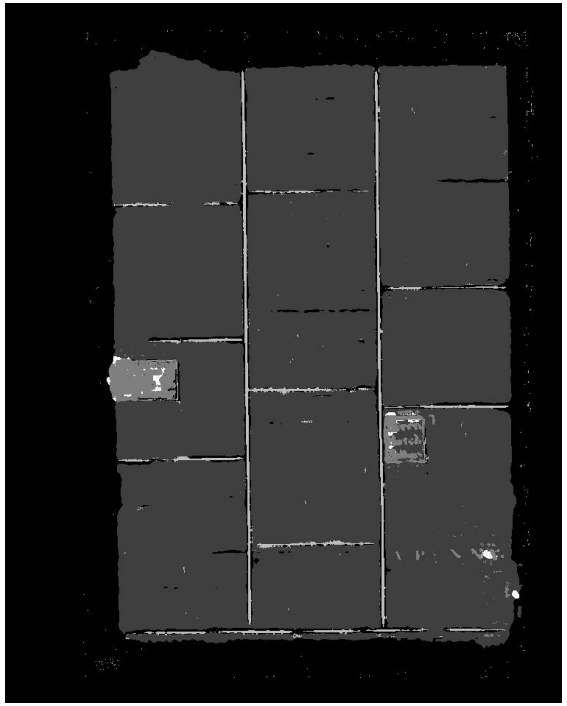
Figure/Graph Extraction from Document | Preliminary Results

- Transfer parameters from pre-trained ResNeXt101 64x4d
- Trained on ENP dataset

Document Image



Ground truth

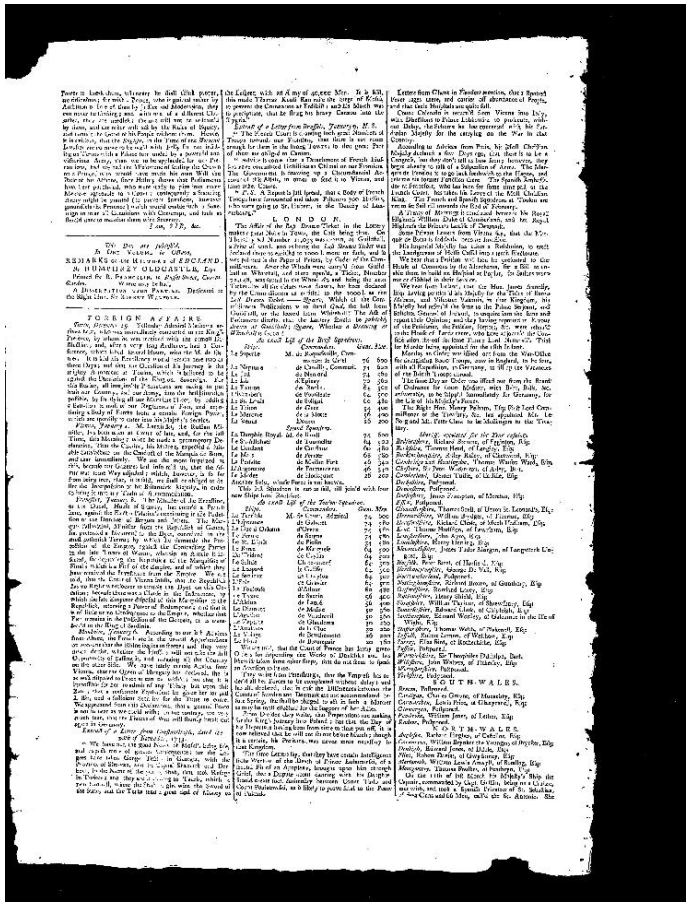


Prediction

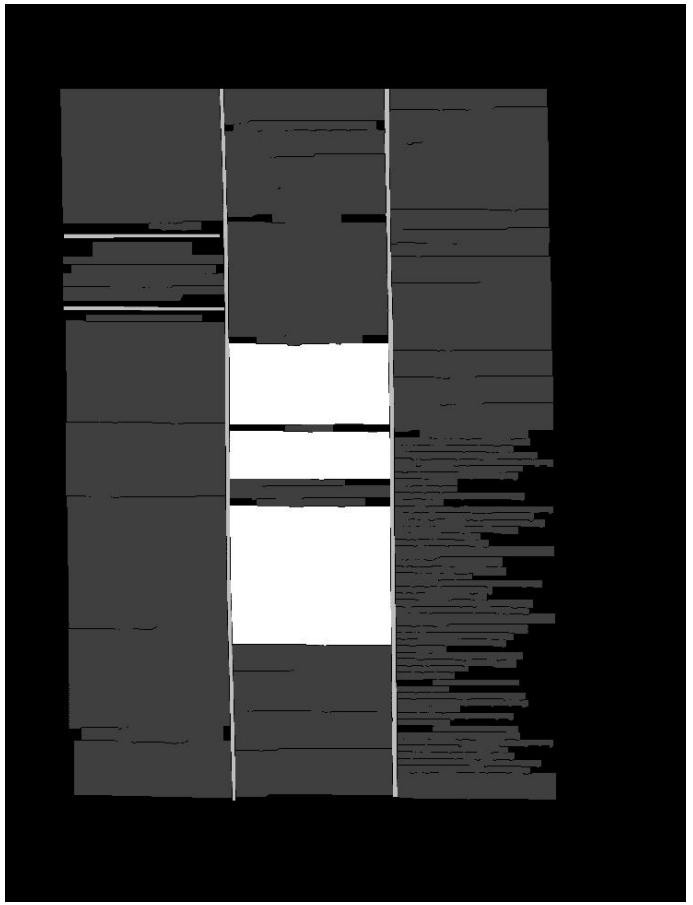
Figure/Graph Extraction from Document | Conclusions

- Promising preliminary results
- Potential applications
  - Segmentation based on content type to increase item-level accessibility
  - Retrieval of figures/graphs for further study
- Challenges
  - U-NeXt still needs more iterations of training
  - Preliminary training indicates that tables may be the hardest type to extract

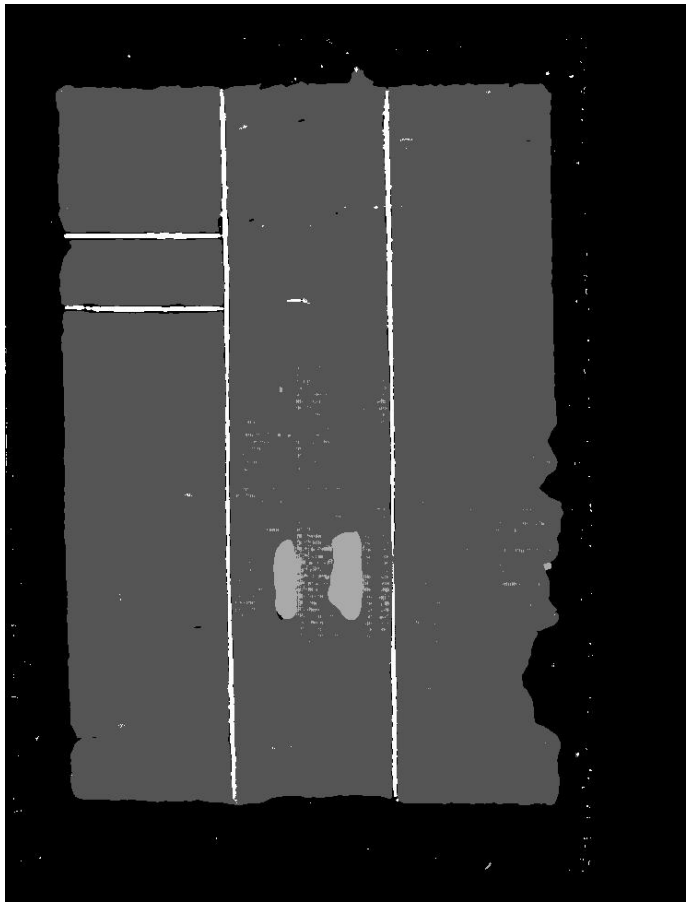
Figure/Graph Extraction from Document | Preliminary Results



Document Image



Ground truth



Prediction

Project 3.2: Text Extraction from Figure/Graph

Objectives | Extract texts from figure/graph

Applications | Metadata generation, OCR for figure/graph caption

Aida

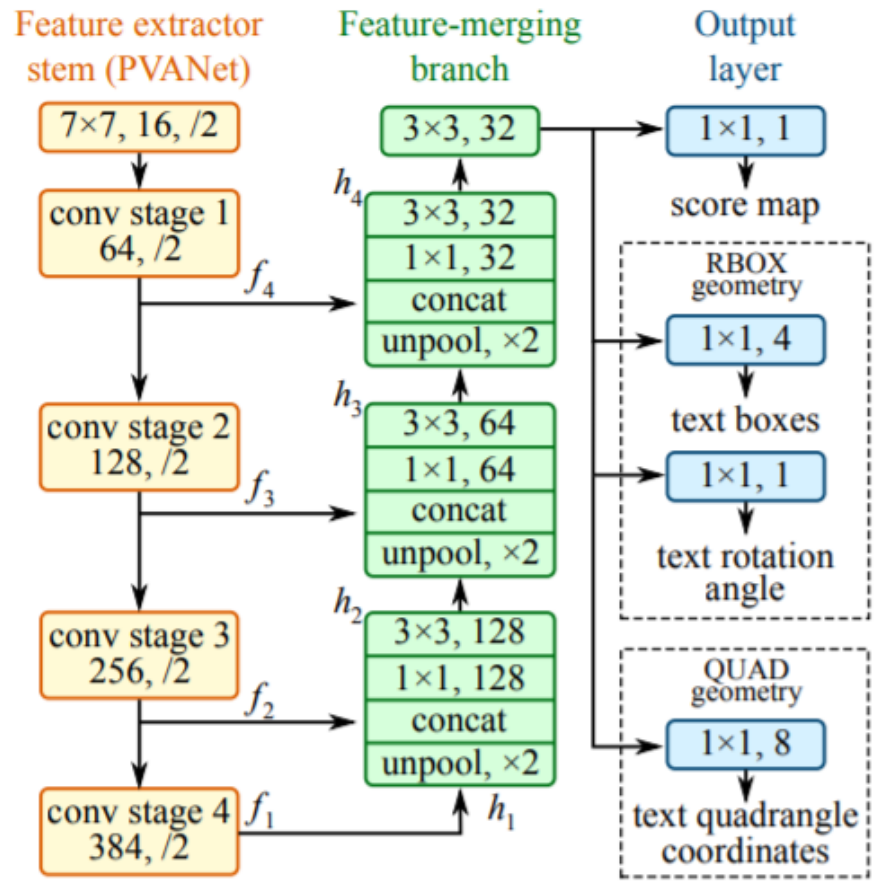
Text Extraction from Figure/Graph | Technical Details

EAST text detector

- EAST: Efficient and Accurate Scene Text detector
- HyperNet + U-Net
- Detect texts in graphic images in any direction

Why applicable?

- figures/illustrations are snippets of a graphic region



Aida

Text Extraction from Figure/Graph | Preliminary Results

Detected Texts

- Performance on detecting texts in newspaper figure/graph is good
- Texts location is recorded

Text Lines

- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

Text Extraction from Figure/Graph | Conclusions

- Promising preliminary results
- Potential application
  - Perform OCR on detected text regions for higher accuracy
  - Extract OCR-ed words in detected text regions as metadata

Project 4.1: Subjective Quality Assessment

Objectives | Access document images based on human perception

Applications | Providing metadata based on human visual perception

Subjective Quality Assessment | Proposal

- Adding an interface to allow users to classify the quality of document images
  - No need for **verbal annotation**
- A simple interface with
  - A drop box having five-level rating scores for **MOS** (i.e., **5-Excellent, 4-Good, 3-Fair, 2-Poor, and 1-Bad**)
  - Buttons, if detailed aspects such as contrast, range-effect, background-cleanness, and content density are needed

Subjective Quality Assessment | Benefits

- A human perception-based document image quality assessment (DIQA) database that can support ***further studies and experiments*** such as machine learning model training
- A ***publicly available*** database can draw attention to more research teams for research competition in academia
- Trained machine learning mode could ***enhance the filter or query search*** in the new UI of Beyond Word to sort images based on their quality

Project 4.2: Objective Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People

Applications | Providing quality scores for machine reading on four criteria: (1) *skewness*, (2) *contrast*, (3) *range-effect*, and (4) *bleed-through*

Aida

Objective Quality Assessment | Technical Details

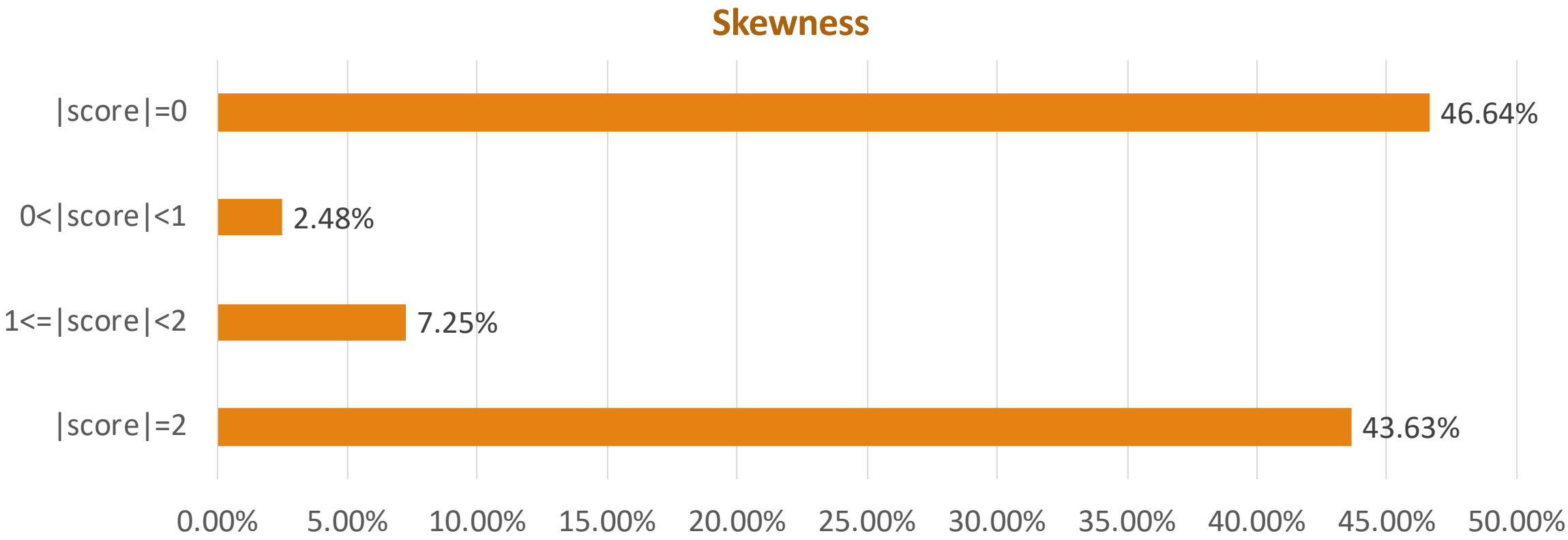
- Objective quality assessment on four criteria
  - *Skewness, Contrast, Range-effect, Bleed-through*
  - *Based on the DIQA programs developed at Aida @ UNL (previously tested using Chronicling America's repository of archived newspaper pages*
  - *Not* directly machine learning related
- **Why**?
  - Help identify images that need pre-processing
  - Reduce unnecessary workload for pre-processing images
  - Indicate general qualities of the dataset

Aida

Objective Quality Assessment | Datasets

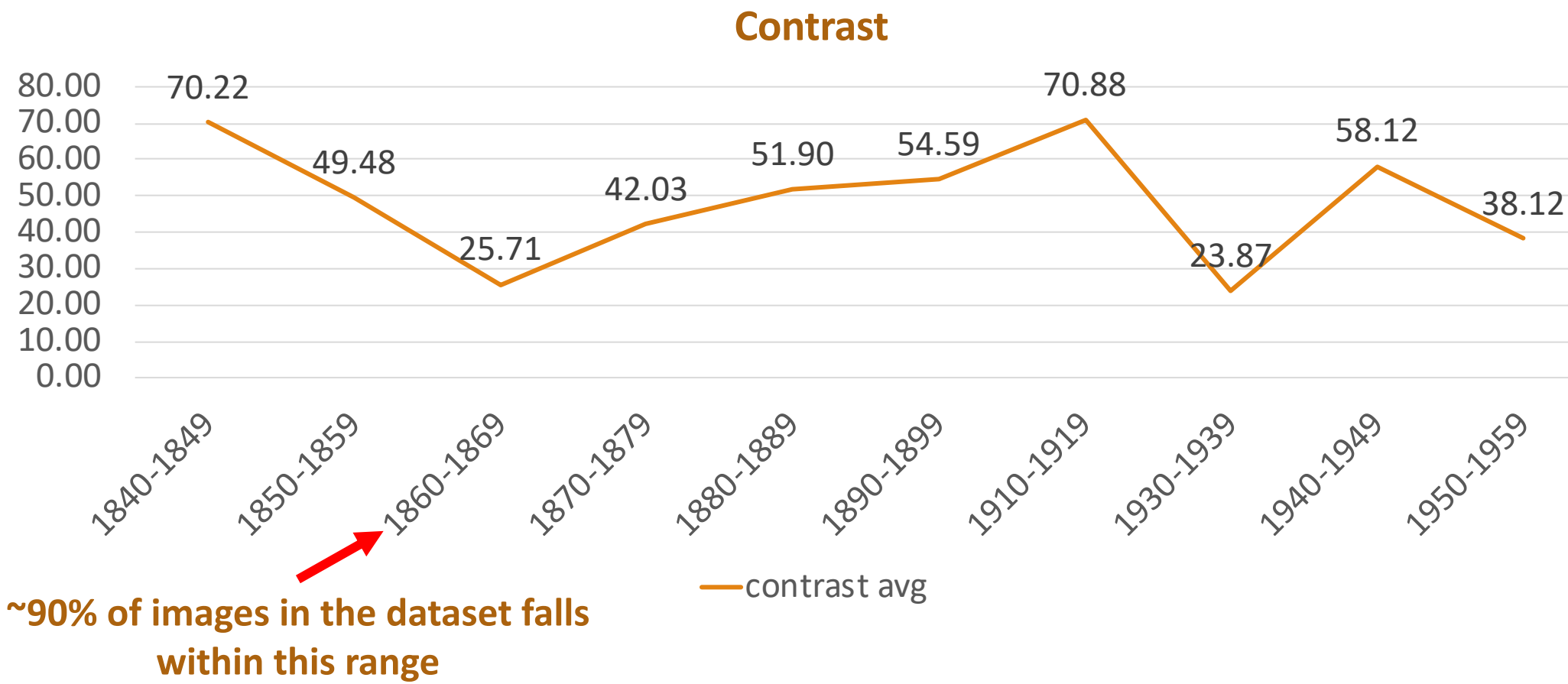
- The Civil War collection within By the People:
  - 36003 images were downloaded
  - 35990 images passed the DIQA program
    - *13 images failed as they barely had texts (see examples later)*

Objective Quality Assessment | Experimental Results

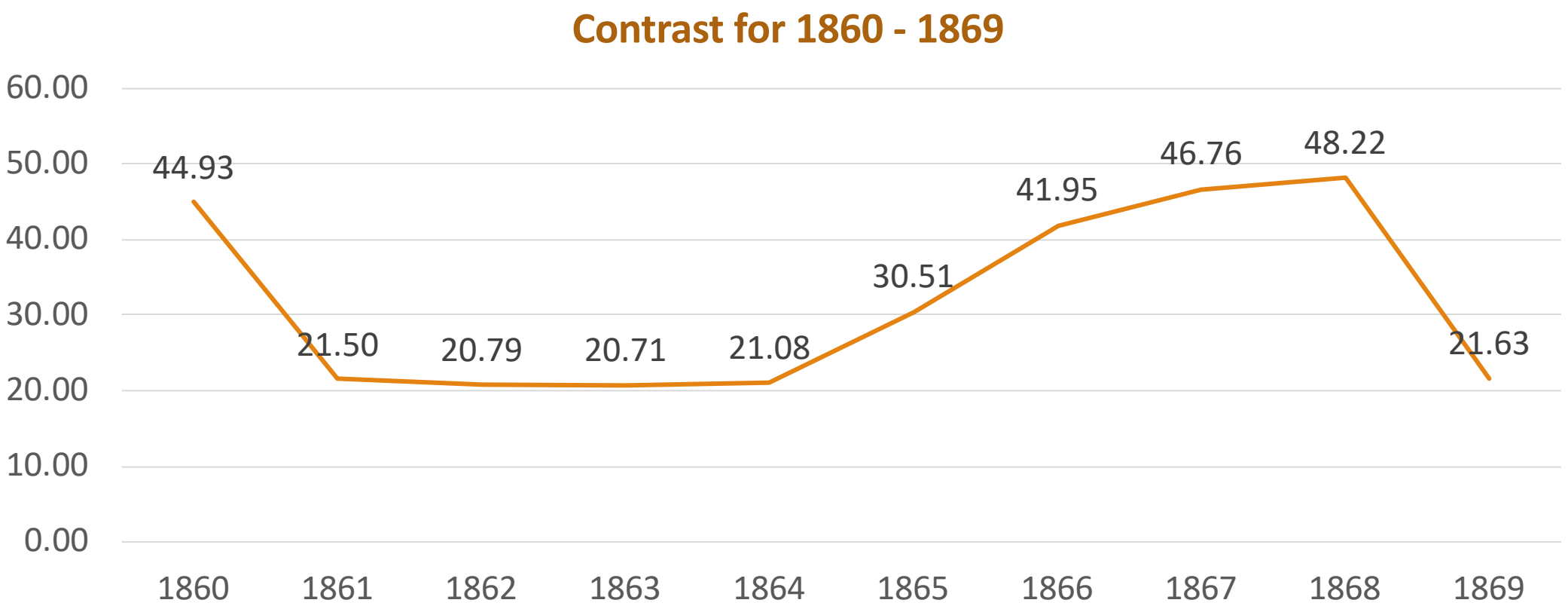


Aida

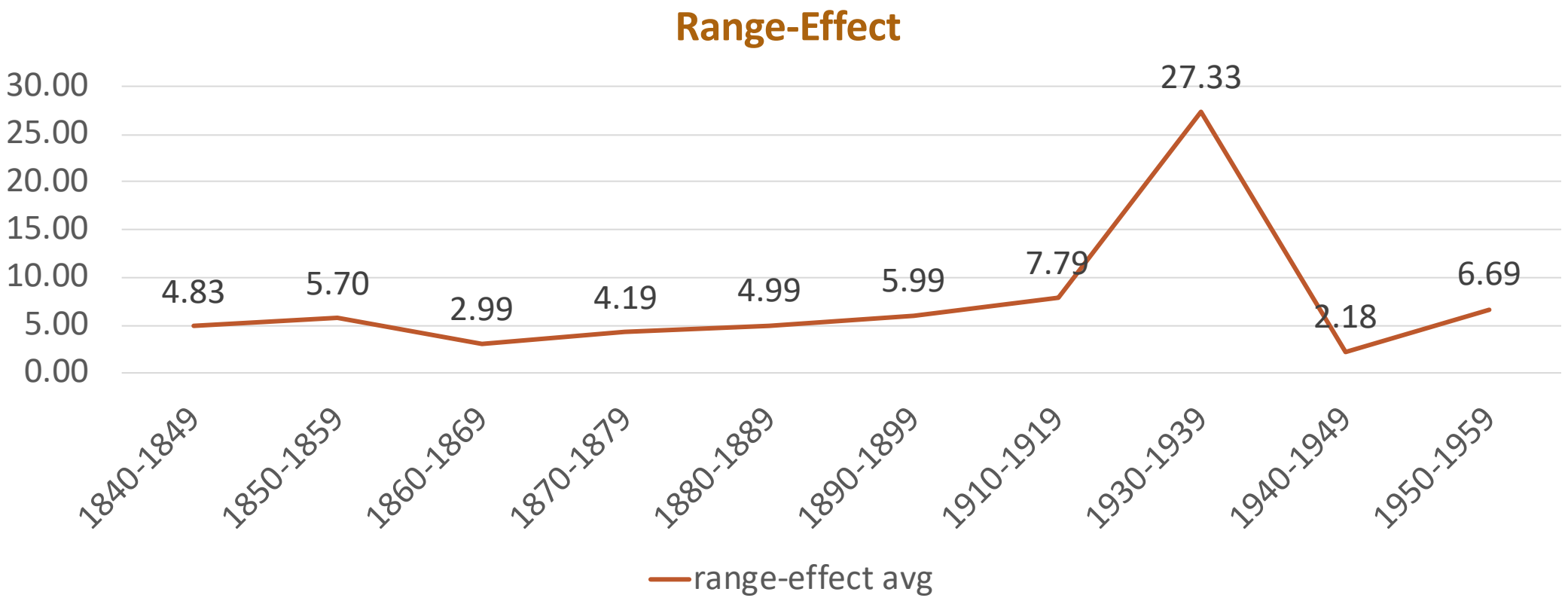
Objective Quality Assessment | Experimental Results



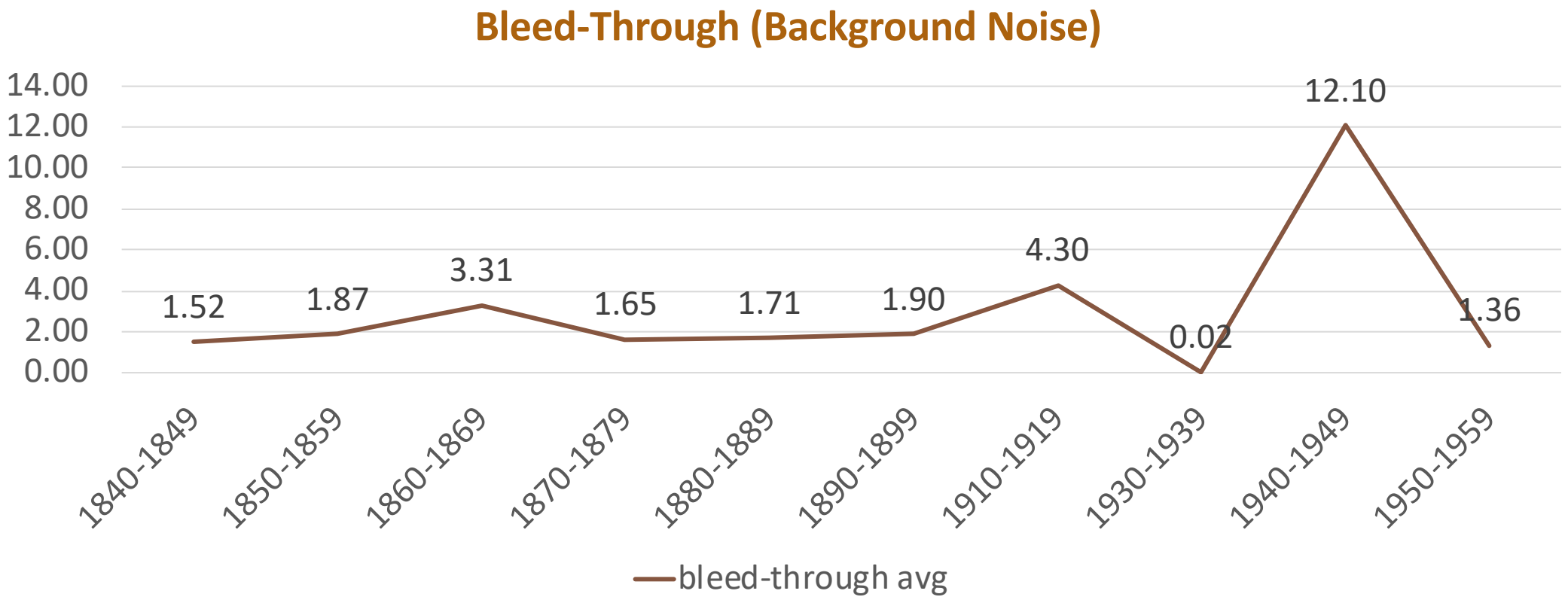
Objective Quality Assessment | Experimental Results



Objective Quality Assessment | Experimental Results



Objective Quality Assessment | Experimental Results

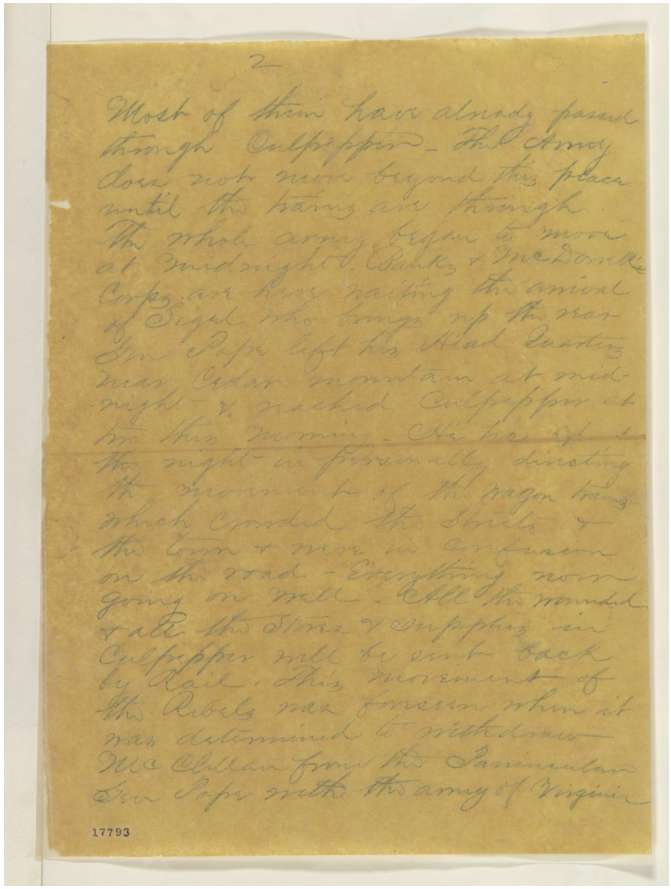


Objective Quality Assessment | Observations

- Must say something about your assessment. Good? Bad? What about the images?

Objective Quality Assessment | Potential Issues

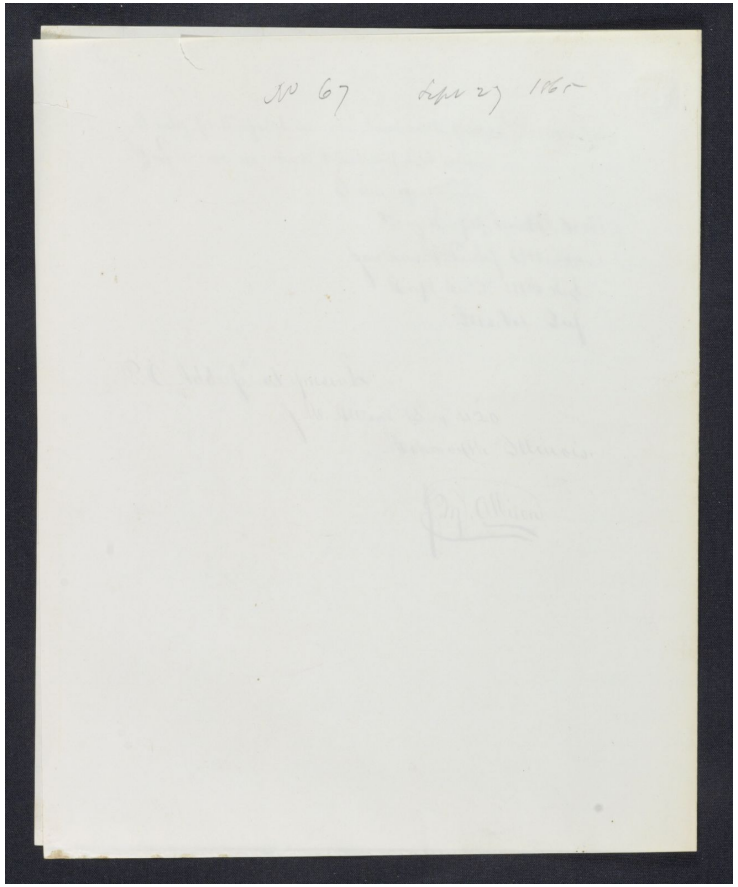
- Numerous images with yellowish background and faded inks
- They are hard to read even to human eye
  - Contrast could be lowered
  - Skewness could be almost impossible to compute



Aida

Objective Quality Assessment | Potential Issues

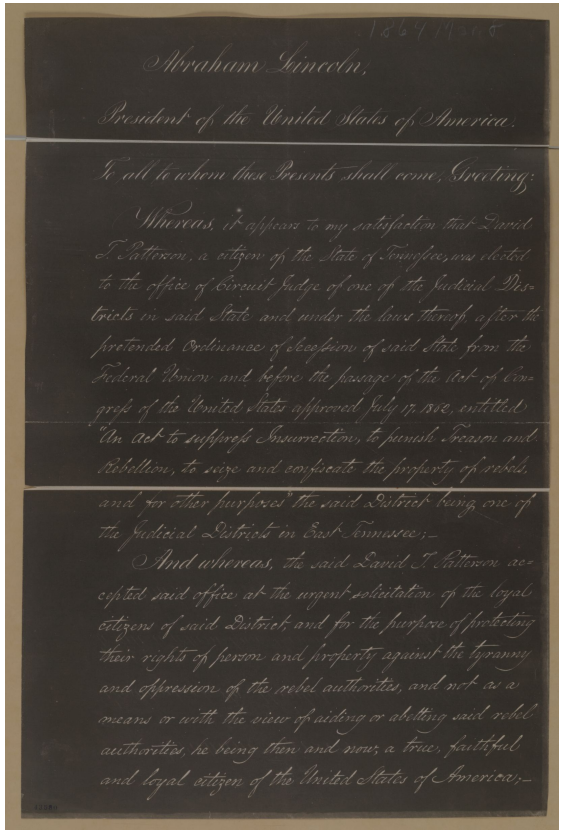
- Numerous images are covers or labels of a series
- These images are largely blank
  - Contrast is poor
  - Histogram equalization might be able to enhance the quality



Aida

Objective Quality Assessment | Potential Issues

- There are color-inverted images from microfilm
  - Renders bleed-through assessment useless



Aida

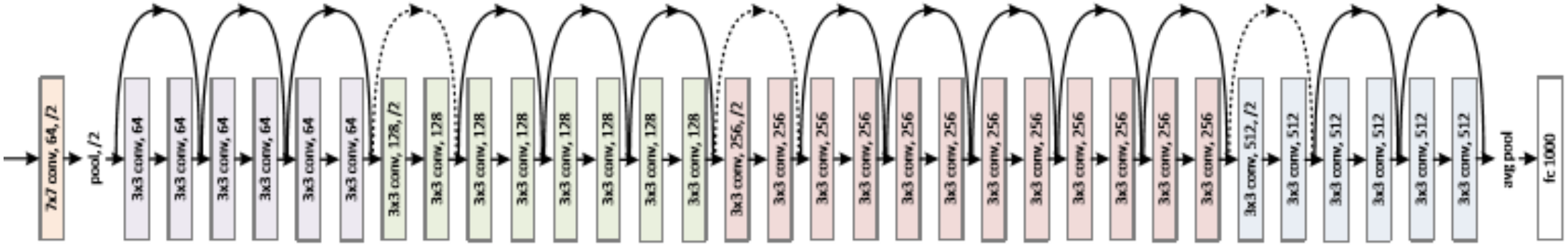
Project 5: Digitization Type Differentiation: Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*

Applications | Metadata generation, pre-processing policy selection

Digitization Type Differentiation | Technical Details

- Pre-trained ResNeXt is adopted
- Attached output layers are two dense layers with a 1D output vector
- **The pre-trained ResNeXt can classify images to 1000 different categories**
- The pre-trained ResNeXt is a good feature extractor
  - Number of parameters: 94.1 million → 12.6 million

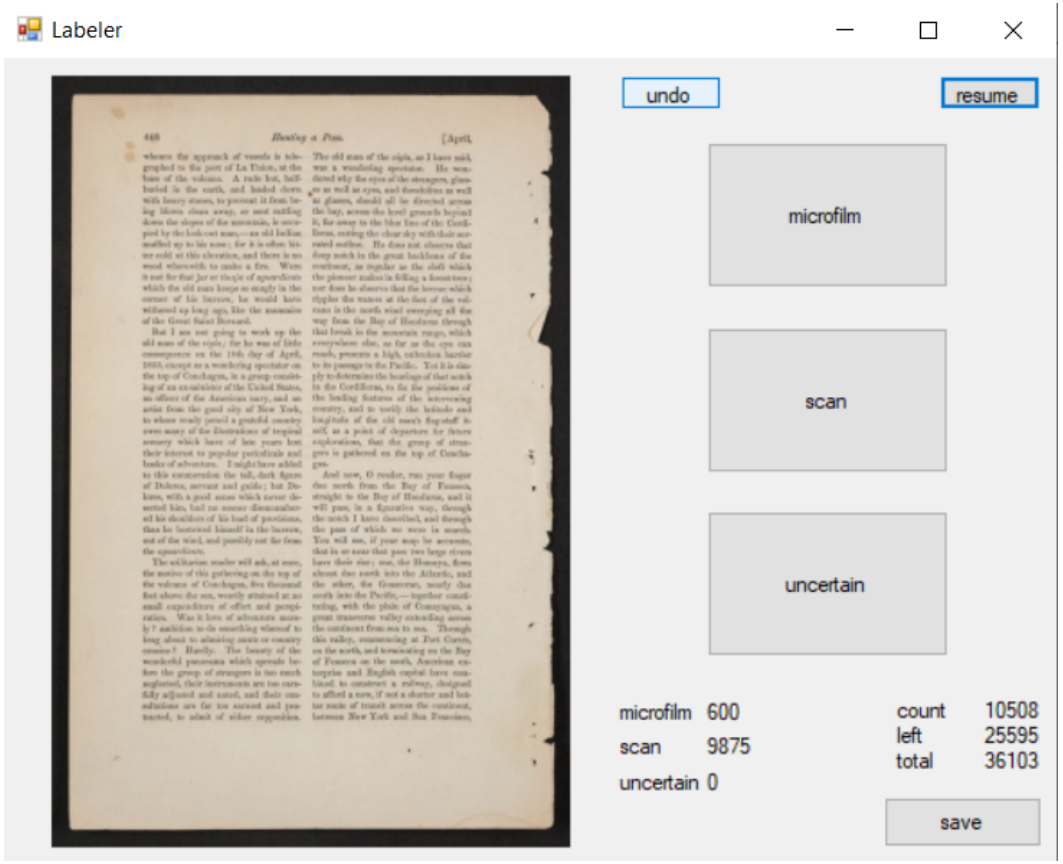


Aida

Digitization Type Differentiation | Datasets

- Created from the Civil War collection within By the People
- A manually created database by ***randomly*** choosing 600 images on scanned materials and 600 images on microfilm materials
- The randomization was performed by shuffling the entire list of 36,003 images in the collection
- The randomization ensured that images in the collection have a fair chance to be chosen
- The randomization seed was fixed to ensure the experiments can be reproduced

Digitization Type Differentiation | Datasets

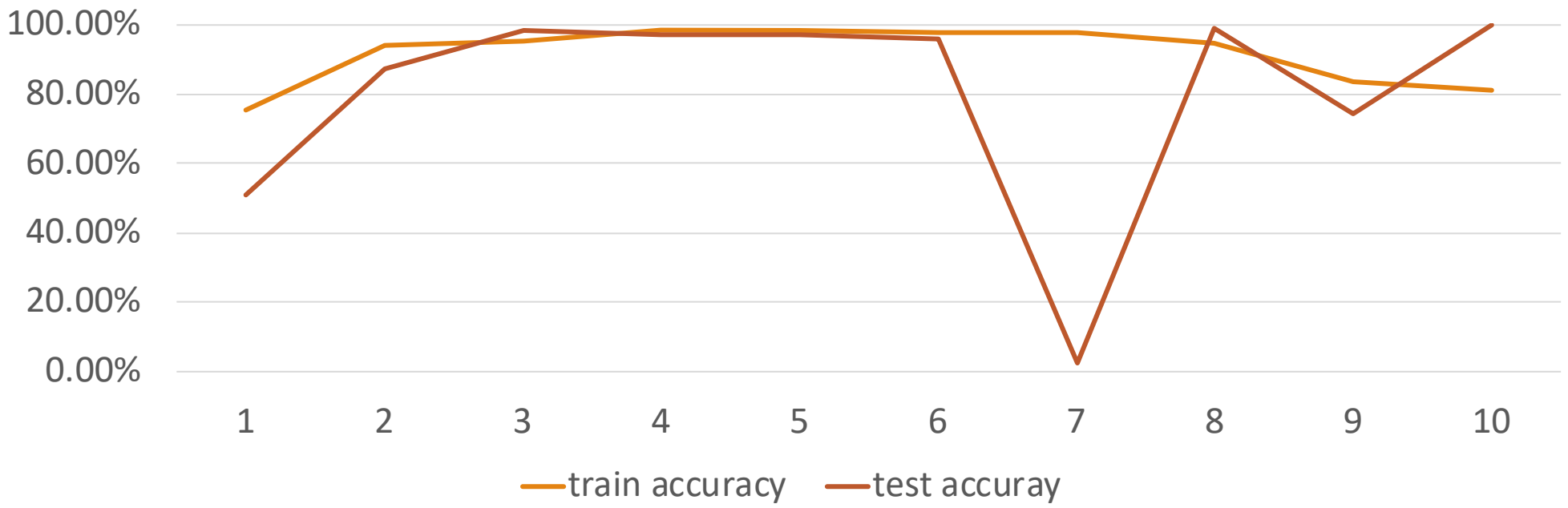


- **Rough estimate**: Based on 10,508 images that was processed, ***ratio of images from microfilm to scanned materials is about 1:16***

Aida

Digitization Type Differentiation | Experimental Results

- With pre-trained ResNeXt,
  - It only took **one** iteration to reach more than 90% accuracy on training set, and
  - It only took **two** iterations to reach more than 90% accuracy on testing set



Digitization Type Differentiation | Experimental Results

- The best test iteration result was able to 100% correctly classify all images

		Ground Truth	
		Scanned	Microfilm
Prediction	Scanned	60	0
	Microfilm	0	60

Aida

Digitization Type Differentiation | Conclusions

- Existing pre-trained model can be easily extended to more designated tasks
- The extended model only need a small set of labeled data to reach near-perfect performance in this task
- Automated digitization type differentiation is ***readily*** achievable.

Digitization Type Differentiation | Tips on Choosing …

- **How** to choose pre-trained model from the “zoo” (or the ”kitchen”)?

Task Type	Model
Type differentiation/classification, with limited computing power	Mobile Net
Type differentiation/classification, with fair amount of computing power	ResNet, ResNeXt
Type differentiation/classification, with good amount of computing power	VGG Network, Inception
Task needs to locate or extract object/figure/graph, based on the amount of computing power	Combine a U-shaped network
Task needs to refine extracted location, and locations may be overlapped	HyperNet

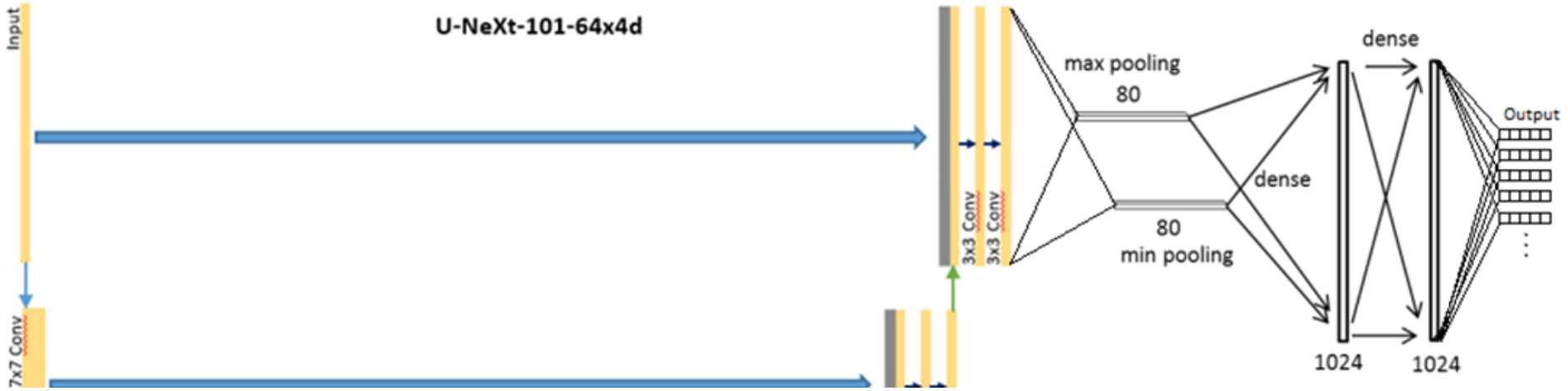
Questions ?

Thank you very much for your participation.

Thanks to Library of Congress + UNL Collaboratory

Subjective Quality Assessment | Technical Details

- Fine tuning pre-trained U-NeXt in Project 1
- **Difference:** DIQA need only high-level score on image quality
  - Instead of 2D matrix output, subjective quality assessment only need 1D vector
  - Elements of the 1D output are image quality scores, such as Mean Opinion Score



Subjective Quality Assessment | Datasets

- Machine Learning, especially for deep learning, requires large amounts of labeled data for training
- Current existing quality assessment databases contain only quality scores for machine perception
  - Previous Aida @ UNL work: Document Image Quality Assessment (DIQA) for Chronicling America newspapers
- **Challenge**
  - Lack of human perception-based DIQA database